Bias Remediation in Driver Drowsiness Detection Systems Using Generative Adversarial Networks

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 21, 2020, accepted March 8, 2020, date of publication March 19, 2020, date of current version March 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981912

Bias Remediation in Driver Drowsiness Detection Systems Using Generative Adversarial Networks

MKHUSELI NGXANDE[1], JULES-RAYMOND TAPAMO[1], (Member, IEEE), AND MICHAEL BURKE[2], (Member, IEEE)

[1]School of Engineering, University of Kwa-Zulu Natal, Durban 4041, South Africa
[2]School of Informatics, Institute of Perception, Action and Behaviour, University of Edinburgh, Edinburgh EH8 9AB, U.K.

Corresponding author: Mkhuseli Ngxande (mngxande@gmail.com)

This work was supported by the University of Kwa-Zulu Natal.

ABSTRACT Datasets are crucial when training a deep neural network. When datasets are unrepresentative, trained models are prone to bias because they are unable to generalise to real world settings. This is particularly problematic for models trained in specific cultural contexts, which may not represent a wide range of races, and thus fail to generalise. This is a particular challenge for driver drowsiness detection, where many publicly available datasets are unrepresentative as they cover only certain ethnicity groups. Traditional augmentation methods are unable to improve a model's performance when tested on other groups with different facial attributes, and it is often challenging to build new, more representative datasets. In this paper, we introduce a novel framework that boosts the performance of detection of drowsiness for different ethnicity groups. Our framework improves Convolutional Neural Network (CNN) trained for prediction by using Generative Adversarial networks (GAN) for targeted data augmentation based on a population bias visualisation strategy that groups faces with similar facial attributes and highlights where the model is failing. A sampling method selects faces where the model is not performing well, which are used to fine-tune the CNN. Experiments show the efficacy of our approach in improving driver drowsiness detection for under represented ethnicity groups. Here, models trained on publicly available datasets are compared with a model trained using the proposed data augmentation strategy. Although developed in the context of driver drowsiness detection, the proposed framework is not limited to the driver drowsiness detection task, but can be applied to other applications.

INDEX TERMS Population bias, GAN, visualisation, CNN.

I. INTRODUCTION

The ability of Artificial Intelligence systems (AI) to automate decision-making capabilities in human daily lives is increasing rapidly. As a result, these systems influence human interaction with the real world and are transforming the future. Their decision-making capabilities typically rely on large training datasets that learn and extract useful patterns in an automated way. Unfortunately, if these systems are trained on datasets that do not have a complete representation of real-world scenarios, they may be prone to bias and prejudice society.

One application where the use of deep learning techniques is increasingly gaining popularity and in which unrepresentative training datasets can lead to negative consequences is that of driver drowsiness detection. One of the primary objectives of the motor industry is passenger safety. Road related accidents are a primary cause of injuries and death among the human population [1]. Among the factors leading to accidents, driving while drowsy is of particular concern. This has prompted the automobile industry to make efforts to develop detection systems that improve driver safety. Monitoring driver behaviour through computer vision and machine learning techniques that detect drowsiness and warn the driver is an increasingly popular technique under investigation by the motor industry. Statistics reveal that a high rate of accidents is caused by drowsy drivers and 20% of serious accidents arise from a failure of driver's judgement and their inability to control the vehicle in the drowsy state [2]. In addition, the World Health Organisation (WHO) reveals that deaths arising from road traffic crashes have increased to 1,35 million in the year 2018 [3]. The report by WHO further

The associate editor coordinating the review of this manuscript and approving it for publication was Fan Zhang.

shows that nearly 3 700 people die on roads every day. This is a particular concern in Africa, which only has 2% of the world's cars, but has the highest accident rate, that accounts for 20% of road deaths [4]. These are alarming findings, which urgently need to be addressed. However, the development of robust driver drowsiness detection systems is still a challenge in both academia and industry.

In the automobile industry, several attempts have been made to monitor driver's state over time by considering various methods such as vehicle-based measures, physiological signals, and behavioural measures. Vehicle-based methods make use of car electronics together with appropriate sensors [5]. These sensors are usually placed on the pedals, steering wheels and often include cameras around the car [6]. Unfortunately, these methods mostly rely on the state of the car and the surrounding environment and focus less on the driver's state. On the other hand, physiological methods monitor the driver's state using physiological signals which can be recorded using devices such as Electroencephalogram (EEG), Electrooculogram (EOG), Electrocardiogram (ECG), or Electromyogram (EMG) [7]–[9]. These devices yield accurate results because the readings measure brain activity [10]. However, physiological methods are invasive as they typically require a device to be placed on the driver to record signals. These devices can make the driver uncomfortable and are considered impractical for real-time drowsiness detection systems.

In contrast, behavioural methods are non-invasive methods that make use of a mounted camera to track the face of the driver and measure the level of drowsiness based on facial features. There are numerous facial features that can be used to measure drowsiness from the camera feed including eye state, yawning, and head position [11]–[13]. Behavioural methods can be combined with machine learning techniques to produce more robust systems. A meta-review [14] examined Hidden Markov Models (HMM), CNNs, and Support Vector Machines (SVM) and concluded that CNNs performed better than other techniques, although SVMs were most commonly used. The success of CNNs has been shown in many computer vision tasks including object classification, segmentation, and object tracking [15]–[17]. CNN architectures require a large amount of training data to learn a suitable representation for a given task. Unfortunately, when it comes to driver drowsiness detection, there are a limited number of publicly available training datasets, and some datasets are not published because of security and privacy reasons preventing the publication of people's faces. In addition, publicly available datasets are often unrepresentative as these may not cover a wide variety of ethnicities. For African contexts, this poses a challenge since the population is diverse and individuals can have many different facial attributes. Models trained using publicly available datasets do not generalise well in an African context [18].

The limitations of datasets that fail to cover a wide range of ethnicities lead to bias in trained models when it comes to contexts with different nationalities. Prior work has shown that visualisation techniques can be used to identify bias in training datasets by identifying population groups where a classifier tends to fail [18]. This paper makes the following contributions in addressing population bias in driver drowsiness training datasets:

- Introduces a novel framework that remedies generalisation failures in under represented population groups in the training dataset, which boosts the performance of drowsiness detection across all population groups.
- Introduces a sampling strategy that identifies individuals with facial features where the network is failing.
- Shows how a GAN that generates realistic images can be used to produce training data for those races or individuals where the model is failing.

The framework relies on two primary components, population bias visualisation and a GAN. The GAN generates realistic images of individuals (drowsy and awake) in these population groups, which are used for retraining the ResNet model used for drowsiness detection with new parameters i.e learning rate and epoch sizes to reduce overfitting and improve the detection accuracy. The population bias visualisation is used to group races by similarity and identifies where the model is failing to generalise. This process is iteratively repeated until convergence.

This paper is organised as follows. Section II provides an overview and background work, which is followed by a discussion on GANs, population bias and CNN visualisation. Details around our framework are discussed in section III, and with training information and a description of the datasets used in this paper. Experimental details and results are presented in section IV. Finally, Section V provides a brief conclusion of our work.

II. BACKGROUND AND RELATED WORK

In this paper, a novel framework that boosts the performance of CNNs for driver drowsiness detection is presented. This is accomplished by highlighting regions where the model is failing and passing similar GAN generated images to the model for retraining. This strategy is based on boosting, where a weak classifier is iteratively re-weighted to make it a strong classifier. In this section, we explain related works on data augmentation and visualisation strategies.

A. GENERATIVE ADVERSARIAL NETWORKS

Since their introduction in 2014 by Goodfellow *et al.* [19], GANs have shown great success in many computer vision tasks, including pose guided person image generation [20], domain transfer [21], super-resolution [22], and text to image applications [23]. Many variants of GAN architectures have been developed, such as Wasserstein Generative Adversarial Networks (WGAN) [24], Wasserstein Generative Adversarial Networks with Gradient Penalty (WGAN-GP) [25], and Deep Convolutional Generative Adversarial Network (DCGAN) [26]. These architectures have been proposed to improve the original GAN architecture for various tasks.

The original GAN architecture is composed of two neural networks namely, a generator G and a discriminator D which are trained by playing a mini-max game against one another. In the case of data augmentation, we utilize the domain shift of one image to another domain. In the transfer of awake states to drowsy states, where we seek to learn a generator distribution p_g over data x, the generator creates a mapping function, parameterized by θ_g from a prior latent distribution $p_z(z)$ to data space $G(z;\theta_g)$. The discriminator $D(x;\theta_d)$, on the other hand, learns parameters θ_d to distinguish whether images are from the training data or from the generator. The mini-max game function $V(G, D)$ is expressed as follows:

$$\begin{aligned} min_G\, max_D\, V(D,G) &= \mathbb{E}_D + \mathbb{E}_G \\ \text{where } \mathbb{E}_D &= \mathbb{E}_{x\backsim p_{data}(x)}[\log D(x)] \\ \mathbb{E}_G &= \mathbb{E}_{z\backsim p_z(z)}[\log(1 - D(G(z)))] \end{aligned} \tag{1}$$

Unfortunately, the original architecture is limited in that there is no flexibility in generating desired outputs. To overcome this problem, conditional GANs were introduced as an extension that introduces additional information to both the networks [27]. This additional information allows the flexibility of producing controllable outputs from the training dataset. The additional information, y, is typically a label which is applied to the resulting image, for example in our case this is the awake or sleepy state. The mini-max objective function from equation (1) is then updated as follows:

$$\begin{aligned} min_G\, max_D\, V(D,G) = \mathbb{E}_{x\backsim p_{data}(x)}[\log D(x|y)] \\ + \mathbb{E}_{z\backsim p_z}(z)[\log(1 - D(G(z|y)))] \end{aligned} \tag{2}$$

In this paper, a controllable GAN is used as a data augmentation technique to balance the training dataset for a driver drowsiness detection task. Data augmentation is a technique that increases the size of a training dataset to reduce the chances of overfitting by the network. In computer vision, the most common way to perform data augmentation is by applying parameterised transformations including random cropping, rotation, scaling, and jittering. In the case of driver drowsiness detection, many available datasets are unrepresentative as these are often captured in specific cultural contexts. In addition, there is a distinct lack of datasets captured in African contexts [33]. Applying common data augmentation strategies to the training dataset improves the results by a small amount. However, standard augmentation is severely limited and is unable to generalise to more complex domain shift problems.

There is much work using GAN architectures for data augmentation. Gupta [28] used conditional GANs for sentiment classification, obtaining a significant improvement against a baseline model that was only trained on real data. The conditional GAN was trained using different strategies including pre-training and noise injection on the training data. Mok and Chung [29] proposed an automatic data augmentation that enables machine learning methods to learn from the available annotated samples efficiently. Their architecture consists of a coarse-to-fine generator which captures the manifold of the training sets. Their proposed method was used on Magnetic Resonance Imaging (MRI) images and achieved improvements of about 3.5% over the traditional augmentation approaches that were compared against. In addition, Wu *et al.* [30] used a multi-scale class conditional GAN to perform contextual in-filling to synthesize lesions onto healthy screening mammograms. For experimentation, three classifications were compared and their method substantially outperformed a baseline model. Jangid [31] used a conditional GAN to augment data in another domain. They named their architecture Data Augmentation Generative Adversarial Networks (DAGAN), which can be trained for low-data tasks using standard stochastic gradient descent approaches. It is clear that GANs can be used as a substitute for traditional augmentation techniques and are particularly valuable where more sophisticated augmentation strategies are required.

B. POPULATION BIAS

Racial bias is a problem that has been raised in the computer vision community, with a specific focus on how to develop machine learning models that guarantee fairness in all ethnicity groups. This problem has been identified as a result of investigations of fairness in machine learning systems that involves humans. Racial bias has been reported in various areas including criminal justice, employment, education, and face recognition systems [34]–[36]. This bias comes from unbalanced training datasets that favor the demographics of the contexts that the application were developed for. As a result, when tested outside of these conditions these algorithms begin to fail. In addition, publicly available datasets often do not capture a wide range of races, while policies of publishing data which contains people's faces prevents some datasets to be published.

Buolamwini and Gebru [37] found that the performance of three commercial gender classification algorithms decreases dramatically for darker-skinned female faces. Moreover, in [38] it is revealed that Amazon's facial recognition tool misidentified photos of 28 US parliamentary members as criminals because of their skin complexion. De-Arteaga *et al.* presented a large scale study of gender bias in occupation classification [39]. Here, a machine learning algorithm learns to classify gender based on first names and pronouns from online biographies. Their study showed that there are gaps when using three different semantic representations such as bag-of-words, Deep Recurrent Neural Networks (DRNN), and word embedding [39]. Benthall and Haynes have investigated supervised learning algorithms and revealed that they are exposed to racial bias because of the differentiation that is embedded in systematic patterns [40].

Abiteboul [41] investigated issues in ethical data management, where he considered bias and violation of data privacy in data analysis. His work discusses factors to be considered when working with data such as fairness, transparency, neutrality, and diversity. To overcome the challenges this introduces, he has proposed a unsupervised learning technique that dynamically detects patterns of segregation in order to

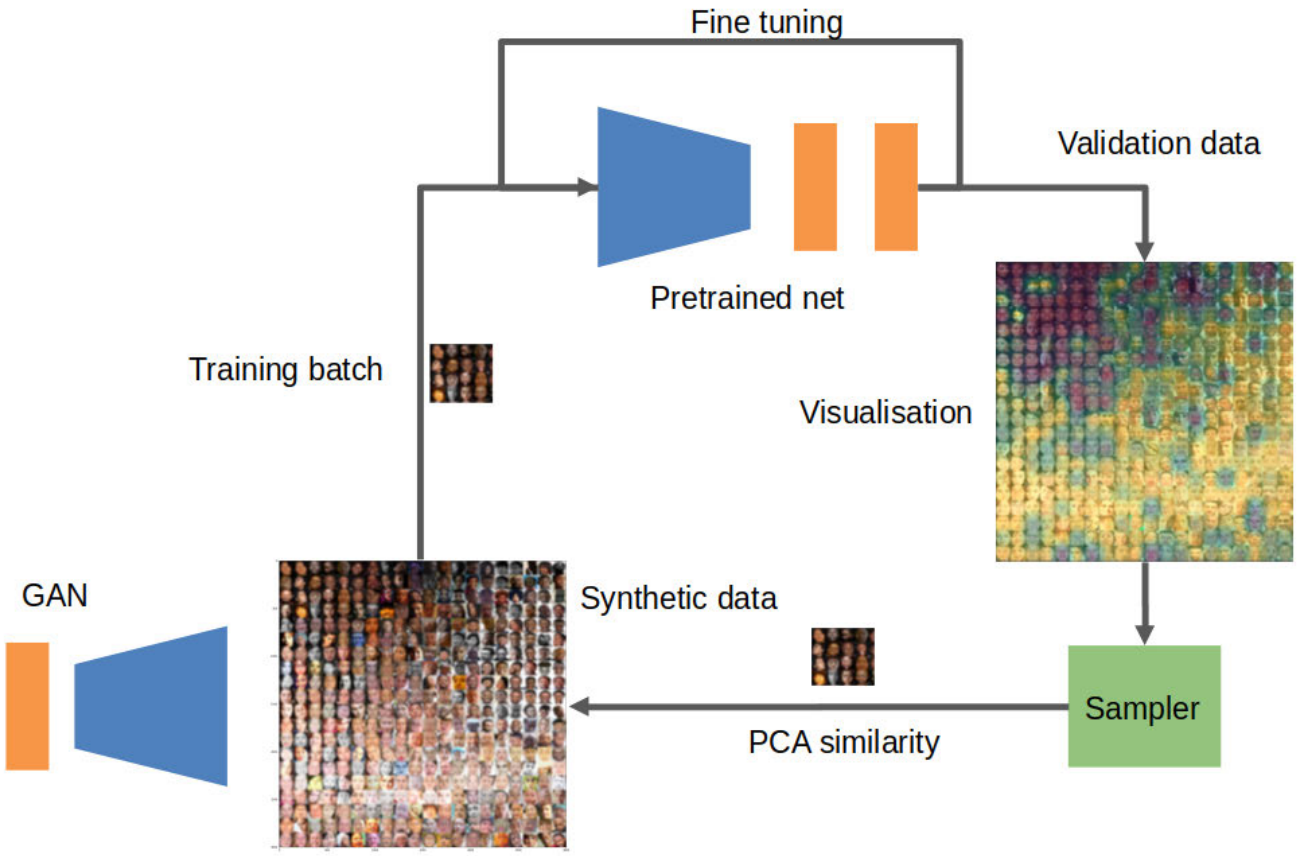


FIGURE 1. The figure shows the proposed bias remediation framework. The first step is the fine tuning of the pre-trained CNN to detect driver state. A population bias visualisation is applied to the testing dataset and highlights where the model is failing to generalise. Images are then sampled where the model is not generalising, and are used to find similar images from the GAN generated images to continue training the model.

mitigate the root causes of social disparities and other factors that can lead to biased models.

In this paper, we seek to address the racial bias introduced by imbalanced training sets, by generating images of awake and drowsy people that look like those for whom the model is failing, so as to improve generalisation.

III. PROPOSED FRAMEWORK FOR BIAS REMEDIATION

This section discusses our framework. Figure 1 illustrates the general architecture of the framework. A pre-trained Resnet model is used for classifying the driver's state, after fine-tuning the final layers with fully connected layers and binary classification layer. The framework boosts the performance of the Resnet classification model on population groups where it is failing (in our case this is darker skinned individuals).

The framework is composed of four primary components. Firstly, a GAN architecture produces synthetic images of individuals with facial attributes that can be used when retraining the network. The second component is the CNN architecture that predicts the state of the driver, while the third component is a population bias visualiser that highlights regions where the model is performing well and where it is failing to generalise. Lastly, the sampler targets images where the model is not performing well and searches through the synthetic images to find more similar images to those, which are used for model retraining.

GAN architecture - We adopt the architecture for our generative network from Choi *et al.* [42], who have shown impressive results for generating realistic synthetic images in different domains. Our architecture is a conditional GAN that is conditioned to translate facial expression attributes such as awake or drowsy across multiple ethnicity groups. This translation of attributes helps in improving the detection model by supplying images with appropriate features where the drowsiness detector fails to generalise. The generator was modified by replacing the standard convolutions by depth-wise separable convolutions [43]. The benefit of this is to have fewer trainable parameters, while retaining the performance of the network. Furthermore, the generator consists of a stride size of two for down-sampling and 11 depthwise separable convolutions. We used instance normalisation [44] instead of batch normalisation for the generator. For the discriminator network, standard convolutions were retained because the discriminator acts as a classifier and requires greater capacity in order to distinguish between real and fake images. In addition, PatchGANs [45] were adapted for discriminator network because they make use of a fixed-size patch discriminator that is easily applied to 256×256 images.

ResNet architecture - We used a pre-trained ResNet model which comprises 50 layers and was originally trained on CIFAR-10, ILSVRC and COCO2 datasets [46]. We fine-tuned the pre-trained model on the last layers, but modified the prediction function to a sigmoid for binary drowsiness classification.

Population bias visualisation - The population bias visualisation component relies on PCA [47] for dimension reduction. This is followed by sorting the images by similarity and overlaying the prediction error to visualise where the model is failing. Image features are extracted by projecting validation images onto a 2-dimensional grid using PCA. The validation dataset is transformed into an orthogonal subspace where axes (Principal Components) align with the directions of maximum variance in the data. Here, a matrix of images is formed by reshaping images x_i into row vectors (where $i = 1 \ldots N$, and N is the number of images in the dataset) and stacking these vertically to form an N x P matrix. The number of pixels in each image is denoted by P. The PCA data transformation starts by mean centering the matrix of images, which is accomplished by subtracting the mean image from each 1 x P dimensional row vector, X_i in the matrix of images,

$$\hat{\mathbf{X}}_i = X_i - \mu_i \tag{3}$$

where $\mu = (\mu_i, .., \mu_p)$ and $\mu_i = \frac{1}{N}\sum_{i=1}^{N} X_{ij}$ and X_{ij} is the j^{th} pixel of the i^{th}image. The mean centred matrix N x P dimensional matrix of images $\hat{X}$ is then decomposed using singular value decomposition (SVD),

$$\hat{\mathbf{X}} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} \tag{4}$$

Here, $\mathbf{U}$ and $\mathbf{V}$ are $P \times P$ and $N \times N$ orthogonal matrices, respectively. Σ is a diagonal matrix comprising the singular values of $\hat{\mathbf{X}}$ in decreasing order [32]. A reduced dimensional representation of $\hat{\mathbf{X}}$ can be obtained by discarding columns of $\mathbf{U}$ and $\mathbf{V}$,

$$\hat{\mathbf{X}} \approx \mathbf{U}_{0:j}\Sigma_{0:j,0:j}\mathbf{V}^{\mathrm{T}}_{0:j} \tag{5}$$

Here, it means $j + 1$ columns have been retained. As shown above, PCA can project data into a low dimensional coordinate system, with axes provided by the columns of $\mathbf{U}_{0:j}$, and data coordinates given by $\mathbf{V}_{0:j}$. In this work, we retain only two columns ($j = 1$), and project images into a two-dimensional coordinate system. Figure 2 shows the 2D projection (coordinates obtained from $\mathbf{V}_{0:1}$) of facial images in our test dataset. We use this projection to construct a grid of images, grouped by similarity.

We create a uniform coordinate grid and search for the closest image (in the reduced dimensional coordinate system) to each point in the grid. We assign each image a corresponding point, and ensure that no image is duplicated, by removing it from the list of available images once allocated a grid coordinate in order to produce a grid of images that groups individuals by facial similarity, as shown in Figure 3. It is clear that this process successfully groups faces of similar state and complexion together, with darker skinned individuals located towards the top of the image, and lighter skinned individuals towards the bottom. For each image selected, we calculate the error in prediction, to produce a saliency map indicating model quality for the constructed grid of images.

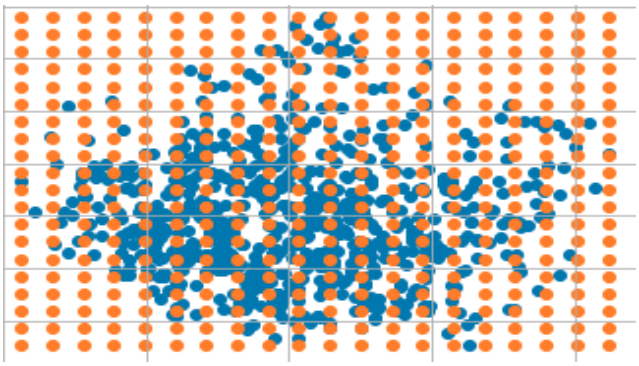


FIGURE 2. The figure shows 2-dimensional grid projection coordinates obtained after applying the linear PCA transformation into 2-dimensional subspace. A uniformly spaced grid is placed over the projected image coordinates, and images are assigned a grid position by finding the closest image coordinate to each grid position, ensuring each image can only be used once. The blue dots represent grid position and the red dots represent PCA projections.

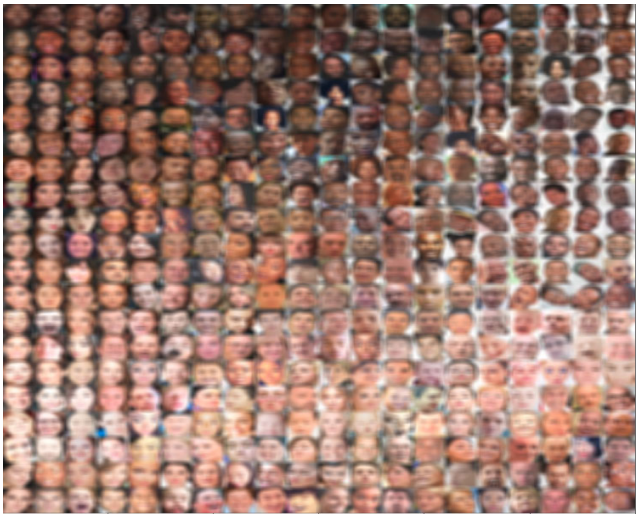


FIGURE 3. The population bias visualisation strategy applies PCA to sort faces by similarity without requiring meta-data.

Image Re-sampling - This step is performed by randomly selecting images according to the error in prediction, termed failure probability here. The failure probability

$$C_i = |y_i - y_t| \tag{6}$$

is calculated by determining the difference between the CNN sigmoid prediction output, y_i and the true label y_t (a binary label encoding drowsy or awake).

This is normalised by the total probability of failure over all N images in the dataset

$$\hat{C}_i = \frac{(1 - C_i)}{\sum_{i=1}^{N}(1 - C_i)} \tag{7}$$

The random selection is performed by categorically sampling images using the weights above. This ensures that images with greater probability of failure are more likely to be sampled. Similar images in the GAN generated dataset are then selected by finding close matching images in the PCA space. The selected images are then used to continue training the ResNet model to increase classification performance.

A. TRAINING

In order to train the GAN architecture, the Adam optimiser [48] was used with $\beta_1 = 0.3$ and $\beta_2 = 0.6$, along with a batch size of 32. We applied a horizontal flipping data augmentation with a probability of 0.5. To train the model, we used a learning rate of 0.0001 for 100 epochs and then linearly decayed the learning rate over the next 100 epochs. This strategy compensates for the fact that our training data is limited. All experiments were carried out on a single Nvidia Tesla K20c GPU, where the training took approximately 12 hours.

Initially, the models were trained on three different datasets (NTHU-drowsy, DROZY, and CEW) and were compared with our framework as illustrated in the results section. The pre-trained ResNet model initially used a learning rate of 0.0001, which was then modified for the rest of the experiments from $1e^{-3}$ to $1e^{-6}$, which was performed for each iteration on our framework. We also used an early stopping strategy to prevent overfitting of the model.

B. DATASETS

This section describes the datasets that were used for training and testing ResNet model. For this work, the NTHU-drowsy, DROZY, and CEW datasets are used.

NTHU-drowsy- was introduced at the 13th Asian Conference on Computer Vision (ACCV2016) [49]. The dataset is split into test and training sets. For training, there are 18 participants (10 men and 8 women) pretending to drive, with 5 scene scenarios for each participant including no-glasses, glasses, glasses at night, no glasses at night, and sunglasses. For evaluation, there are images of 2 men and 2 women. Videos combining drowsy, normal and sleepy states are provided.

DROZY- consists of 14 participants (3 males and 11 females) [50]. Each video is approximately 10 minutes long and is accompanied by the results of psychomotor vigilance tests (PVTs) regarding the drowsiness state. For each participant, the dataset contains a time-synchronized Karolinska Sleepiness Scale (KSS) score [50].

CEW- is a collection of online images of different races (for example Asians and non-Asians with light-skinned faces) and contains about 2423 participants [51]. Among the participants, 1192 have both eyes closed and 1231 have their eyes open. These images were selected from the labeled faces in the wild database.

Validation Dataset- Our validation set contains 1500 faces which some are obtained on the validation sets of the three datasets used. To have a balance and representative validation dataset, we added African faces which we collected for this purpose of drowsiness detection task. These images contain many ethnicity groups with facial drowsiness states.

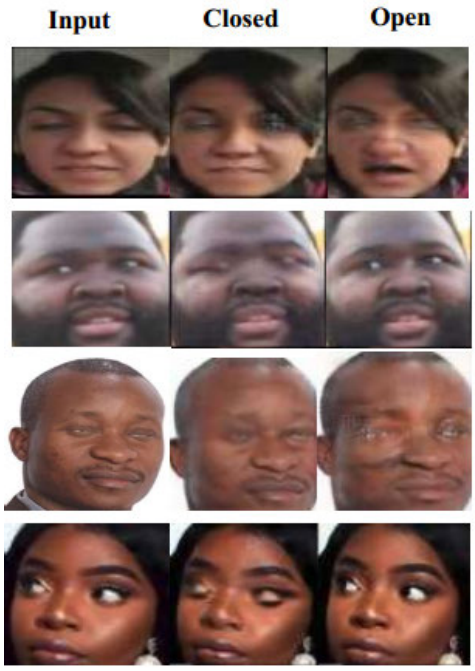


FIGURE 4. Despite being imperfect, examples generated by the GAN serve to improve model performance.

TABLE 1. Model classification accuracy.

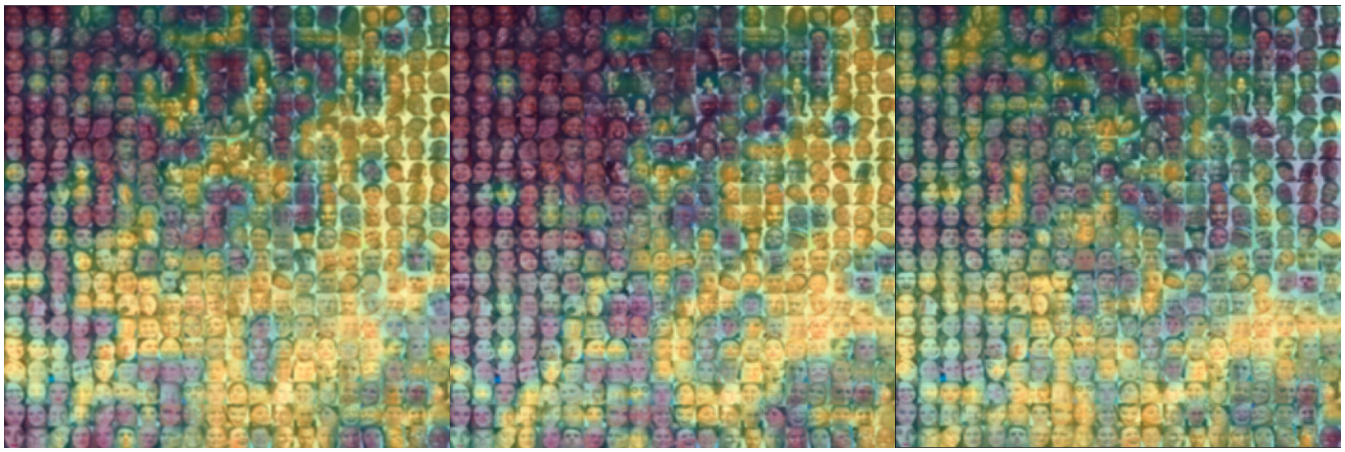
Detection Accuracy				
Iteration	Original Data	GAN Augmentation (validation)	GAN Augmentation with Sampling (validation)	GAN Augmentation with Sampling (test)
1	56.40%	74.70%	87.65%	85.28%
2	55.70%	76.86%	89.87%	85.97%
3	57.08%	77.91%	89.43%	86.42%
4	52.00%	77.01%	90.04%	87.02%
5	59.76%	78.80%	93.66	88.83%
6	60.92%	80.51%	94.5%4	89.05%
7	60.54%	80.93%	96.75%	91.62%

IV. EXPERIMENTAL RESULTS

We first evaluated our proposed framework to results on the publicly available dataset using the pre-trained ResNet models. All the parameters were kept the same for all the first experiments, but thereafter the learning rate and training epochs were modified to prevent overfitting of the models.

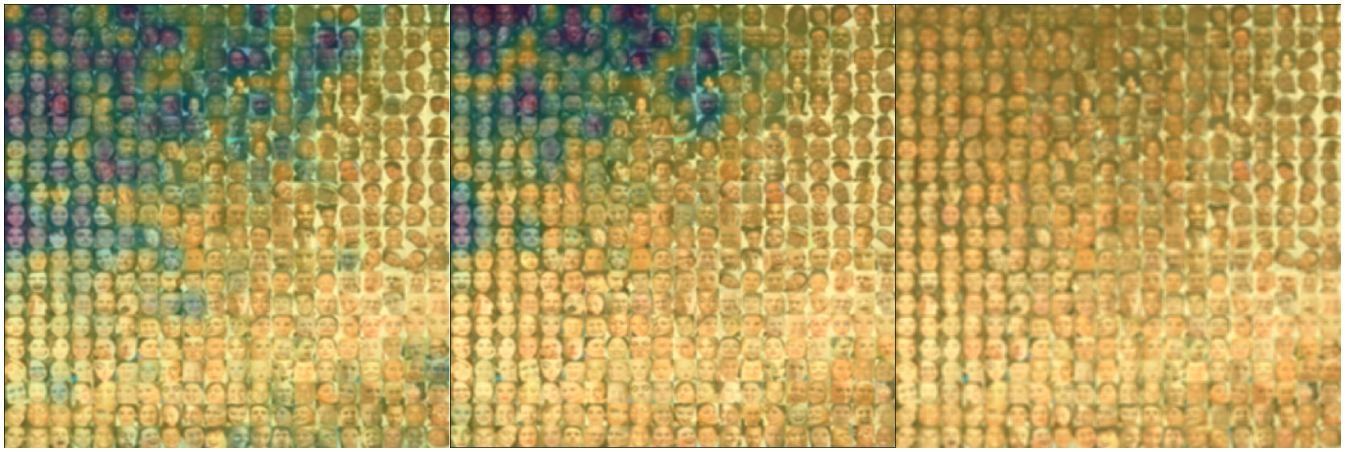
A. GAN AUGMENTATION RESULTS

Figure 4 shows the eye state attribute transfer results on validation dataset. We observed that it is easier to transfer from eyes open to eyes closed. The first row in figure 4 shows that the network was not performing well in transferring from eyes closed to eyes opened, as shown by the blurriness of the image. Despite these limitations, these images still boosted the performance of detecting driver drowsiness task.



(a) ResNet-CEW (b) ResNet-DROZY (c) ResNet-NTHU-drosy

FIGURE 5. The figure shows images produced using the population bias visualisation technique. All the trained models appear to be failing on the population groups on the upper part of the image. The yellow shaded parts indicate where the model performs well, while failures are indicated by the purple shaded parts, which appear mostly on the upper part of the images. The green shaded parts show that the model is also performing well, but with lower probability (0.50 to 0.65).



(a) First visualisation (b) Second visualisation (c) Seventh visualisation

FIGURE 6. The figure shows the progressive improvement using the framework. As the models were fine-tuned and the learning rate was modified, there was improvement and the models reached more certain classification probabilities (0.70 to 0.99). At the seventh cycle we managed to reach optimal performance on the validation set.

B. POPULATION BIAS VISUALISATION RESULTS

Population bias visualisation highlights faces where the model fails to generalise. A re-sampling strategy randomly targets highlighted images where the model is failing and uses similar GAN images for model retraining. In these experiments, we compared our framework to models that are trained on CEW, NTHU-drowsy, and DROZY datasets. These models are then tested on the prepared test set. The GAN synthetic data was generated from the validation set which is prepared to represent a wide variety of races. These synthetic generated images cover a wide variety of races and facial attributes.

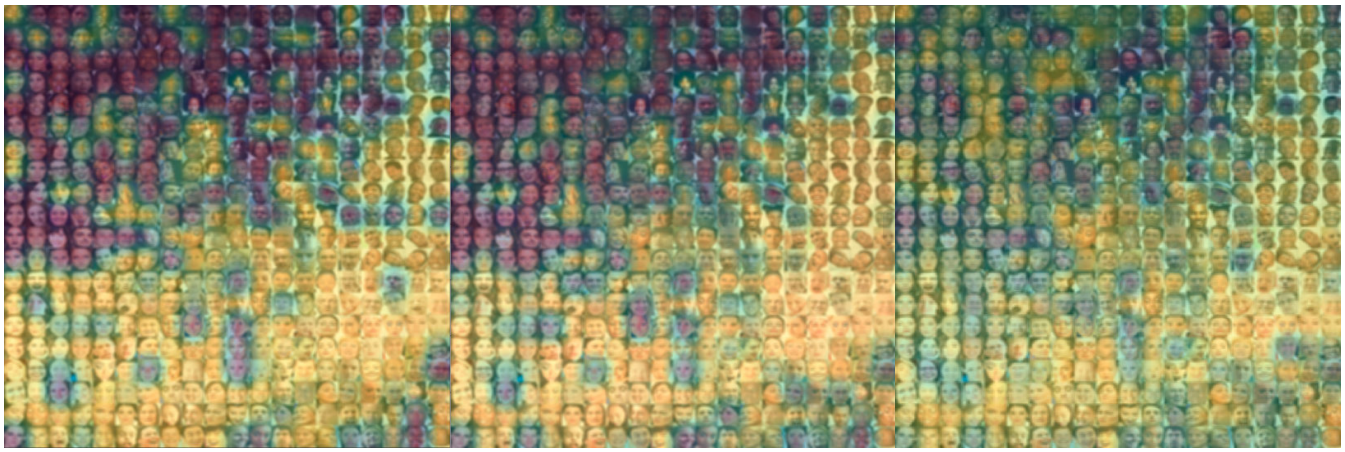
Fig 7 shows baseline experimental results based where we sampled randomly from the augmentation data. It was noted that when we used all the GAN generated images and did not use targeted sampling, model improvement was limited. Targeted sampling is clearly a more effective strategy.

Fig 5 shows the population bias visualisation for the three datasets. It is clear that the models were failing to generalise on darker complexion ethnicity groups, but the proposed framework corrects this. Table 1 shows the accuracy results from the three experiments which show how accuracy is improved with additional iterations of targeted sampling. The table also contains results from previously unseen test images as the model is updated using the sampling technique, which show improvements in line with the validation set.

Fig 6 highlights the improvement of the driver drowsiness detection performance using the proposed framework. Given enough data, the model was able to generalise well to all population groups in the test set.

C. LEARNING RATE RESULTS

Table 2. shows the influence of different learning rates on the performance of the model. A learning rate of $1e^{-3}$ was too high and limited model learning, which results into



(a) First random visualisation (b) Second random visualisation (c) Seventh random visualisation

FIGURE 7. **The figure shows the results when the random selection strategy is used without targeted sampling. We randomly selected GAN generated images and this shows the contribution of the image to the performance of the network.**

TABLE 2. **Learning rate - accuracy tradeoff.**

Learning rates with accuracies	
Learning Rate	Accuracies
$1e^{-3}$	89.70%
$1e^{-4}$	96.30%
$1e^{-5}$	96.91%
$1e^{-6}$	98.01%

overfitting the model. The best performance was observed from $1e^{-4}$ to $1e^{-6}$. The early stopping with learning from $1e^{-4}$ to $1e^{-6}$ strategy was also applied to prevent overfitting which can improve the generalisation on the model. At first, the models were trained with the same number of epochs. It was observed that using the same number of epochs on different learning rates showed traits of overfitting. Therefore, early stopping was used when there is no change in the accuracy.

V. CONCLUSION

In this paper, we introduced a novel framework that can be used to boost the performance of driver drowsiness detection models by reducing bias in the training dataset. Here, a GAN network produces realistic images that are used when retraining a ResNet model on synthetic data generated based on failure cases in validation data.

Faces where the model fails are used to search for similar images in a synthetic dataset produced by the GAN network. These images are used to fine tune a CNN model, and this process is repeated until convergence. A strategy of early stopping with learning from $1e^{-4}$ to $1e^{-6}$ different learning rates for each iteration helped to prevent the chances of overfitting the model.

Importantly, the proposed approach does not rely on any meta-data or assumptions about the race or ethnicity of individuals in the datasets, which is a commonly used approach to determine algorithmic fairness or bias. Requiring this knowledge is potentially problematic as it tends to rely on subjective and controversial racial classifications.

This work has shown that bias in datasets can be addressed to some extent using targeted sampling and generative adversarial networks. However, this process still requires that some training data be available for various population groups, and does not eliminate the need for good, representative datasets. Rather, the proposed approach is intended to remedy more subtle bias introduced by imbalanced datasets, where images of a particular group may be more numerous than that of another.

REFERENCES

[1] *Global Status Report on Road Safety 2018*. Accessed: Apr. 19, 2019. [Online]. Available: https://www.who.int/violence_injury_prevention/road_safety_status/2018/English-Summary-GSRRS2018.pdf

[2] *Driver Fatigue*. Accessed: Feb. 25, 2019. [Online]. Available: http://driverfatigue.co.uk/

[3] *Global Status Report on Road Safety 2018*. Accessed: Mar. 3, 2019. [Online]. Available: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/

[4] *Africa Has 2% of World'S Cars But 20% of Road Deaths'-First Safety Observatory to Curb Horrendous Death Toll*. Accessed: Mar. 5, 2019. [Online]. Available: https://www.wheels24.co.za

[5] W. Wierwille and R. Knipling, "Vehicle-based drowsy driver detection: Current status and future prospects," in *Proc. IVHS Amer.*, Apr. 1994, pp. 245–256.

[6] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: A review," *Sensors*, vol. 12, no. 12, pp. 16937–16953, 2012.

[7] A. Luthra, *Echo Made Easy*. Ahmedabad, Gujarat: JP Medical, 2016.

[8] C. Drewes, "Electromyography: Recording electrical signals from human muscle," Tested Stud. Lab. Teach. Assoc. Biol. Lab. Educ. (ABLE), 2000, pp.248-270, vol. 21.

[9] N. R. Folane and R. M. Autee, "EEG based brain controlled wheelchair for physically challenged people," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 4, no. 6, pp. 2257–2263, 2016.

[10] M. Awais, N. Badruddin, and M. Drieberg, "A hybrid approach to detect driver drowsiness utilizing physiological signals to improve system performance and wearability," *Sensors*, vol. 17, no. 9, p. 1991, 2017.

[11] M. Li and H.-L. Meng, "A method of driver fatigue detection based on multi-features," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 8, no. 10, pp. 107–114, Oct. 2015.

[12] M. V. Rajput and J. W. Bakal, "Execution scheme for driver drowsiness detection using yawning feature," *Int. J. Comput. Appl.*, vol. 62, no. 6, pp. 6–11, 2013.

[13] C. Jacobé de Naurois, C. Bourdin, A. Stratulat, E. Diaz, and J.-L. Vercher, "Detection and prediction of driver drowsiness using artificial neural network models," *Accident Anal. Prevention*, vol. 126, pp. 95–104, May 2019.
[14] M. Ngxande, J.-R. Tapamo, and M. Burke, "Driver drowsiness detection using behavioral measures and machine learning techniques: A review of state-of-art techniques," presented at the Pattern Recognit. Assoc. South Afr. Robot. Mechatronics (PRASA-RobMech), Bloemfontein, South Africa, Nov. 2017.
[15] R. Shima, H. Yunan, O. Fukuda, H. Okumura, K. Arai, and N. Bu, "Object classification with deep convolutional neural network using spatial information," in *Proc. Int. Conf. Intell. Inform. Biomed. Sci. (ICI-IBMS)*, Nov. 2017, pp. 135–139.
[16] S. Jangid and P. S. Bhatnagar, "Semantic image segmentation using deep convolutional neural networks and super-pixels," *Int. J. Appl. Eng. Res.*, vol. 13, no. 20, pp. 14657–14663, 2018.
[17] D. Li and W. Chen, "Object tracking with convolutional neural networks and kernelized correlation filters," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, May 2017, pp. 1039–1044.
[18] M. Ngxande, J.-R. Tapamo, and M. Burke, "Detecting inter-sectional accuracy differences in driver drowsiness detection algorithms," 2019, *arXiv:1904.12631*. [Online]. Available: http://arxiv.org/abs/1904.12631
[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.* 2014, pp. 2672–2680.
[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016, *arXiv:1611.07004*. [Online]. Available: http://arxiv.org/abs/1611.07004
[21] J. Ngiam, D. Peng, V. Vasudevan, S. Kornblith, Q. V. Le, and R. Pang, "Domain adaptive transfer learning with specialist models," 2018, *arXiv:1811.07056*. [Online]. Available: http://arxiv.org/abs/1811.07056
[22] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "ESRGAN: Enhanced super-resolution generative adversarial networks," 2018, *arXiv:1809.00219*. [Online]. Available: http://arxiv.org/abs/1809.00219
[23] C. Bodnar, "Text to image synthesis using generative adversarial networks," 2018, *arXiv:1805.00676*. [Online]. Available: http://arxiv.org/abs/1805.00676
[24] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in *Proc. 34th ICML*, vol. 70, Sydney, NSW, Australia, Aug. 2017, pp. 214–223.
[25] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," 2017, *arXiv:1704.00028*. [Online]. Available: http://arxiv.org/abs/1704.00028
[26] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[27] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784
[28] R. Gupta, "Data augmentation for low resource sentiment analysis using generative adversarial networks," 2019, *arXiv:1902.06818*. [Online]. Available: http://arxiv.org/abs/1902.06818
[29] T. C. W. Mok and A. C. S. Chung, "Learning data augmentation for brain tumor segmentation with Coarse-to-Fine generative adversarial networks," 2018, *arXiv:1805.11291*. [Online]. Available: http://arxiv.org/abs/1805.11291
[30] E. Wu, K. Wu, D. Cox, and W. Lotter, "Conditional infilling GANs for data augmentation in mammogram classification," 2018, *arXiv:1807.08093*. [Online]. Available: http://arxiv.org/abs/1807.08093
[31] A. Antoniou, A. Storkey, and H. Edwards, "Augmenting image classifiers using data augmentation generative adversarial networks," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, Oct. 2018, pp. 594–603.
[32] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM Rev.*, vol. 35, no. 4, pp. 551–566, Dec. 1993, doi: 10.1137/1035134.
[33] J. Parris, M. Wilber, B. Heflin, H. Rara, A. El-barkouky, A. Farag, J. Movellan, M. Castrilón-Santana, J. Lorenzo-Navarro, M. N. Teli, S. Marcel, C. Atanasoaei, and T. E. Boult, "Face and eye detection on hard datasets," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, Oct. 2011, pp. 1–10.
[34] N. Shadowen, "Ethics and bias in machine learning: A technical study of what makes us 'good,'" in *The Transhumanism Handbook*. Cham, Switzerland: Springer, 2019, pp. 247–261.
[35] L. Rhue, *Emotion-Reading Tech Fails the Racial Bias Test*. The Conversation, 2019.
[36] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving Face: Investigating the ethical concerns of facial recognition auditing," 2020, *arXiv:2001.00964*. [Online]. Available: https://arxiv.org/abs/2001.00964
[37] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conf. Fairness, Accountability Transparency*, New York, NY, USA, Feb. 2018, pp. 1–10.
[38] C. Garvie, A. Bedoya, and J. Frankle, *The Perpetual Line-Up. Unregulated Police Face Recognition in America*. Washington, DC, USA: Georgetown Law Center Privacy & Technology, Oct. 2016.
[39] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, S. Geyik, A. Chouldechova, K. Kenthapadi, and A. Tauman Kalai, "Bias in bios: A case study of semantic representation bias in a high-stakes setting," 2019, *arXiv:1901.09451*. [Online]. Available: http://arxiv.org/abs/1901.09451
[40] S. Benthall and B. D. Haynes, "Racial categories in machine learning," in *Proc. Conf. Fairness, Accountability, Transparency*, Atlanta, GA, USA, 2019, pp. 289–298.
[41] S. Abiteboul, "Issues in ethical data management," presented at the 19th Int. Symp. Princ. Pract. Declarative Program., New York, NY, USA, 2017.
[42] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain Image-to-Image translation," 2017, *arXiv:1711.09020*. [Online]. Available: http://arxiv.org/abs/1711.09020
[43] M. Ngxande, J. Tapamo, and M. Burke, "DepthwiseGANs: Fast training generative adversarial networks for realistic image synthesis," in *Proc. Southern Afr. Universities Power Eng. Conf./Robot. Mechatronics/Pattern Recognit. Assoc. South Afr. (SAUPEC/RobMech/PRASA*, Bloemfontein, South Africa, Jan. 2019, pp. 111–116.
[44] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022
[45] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," 2016, *arXiv:1604.04382*. [Online]. Available: http://arxiv.org/abs/1604.04382
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: http://arxiv.org/abs/1512.03385
[47] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," 2015, *arXiv:1511.06448*. [Online]. Available: http://arxiv.org/abs/1511.06448
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[49] *NTHU CVlab-Driver Drowsiness Detection Dataset*. Accessed: Aug. 3, 2018. [Online]. Available: http://cv.cs.nthu.edu.tw/php/callforpaper/datasets/DDD/
[50] Q. Massoz, T. Langohr, C. François, and J. G. Verly, "The ULg multimodality drowsiness database (called DROZY) and examples of use," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–7.
[51] *The Closed Eyes in the Wild (CEW) Dataset*. Accessed: Apr. 19, 2018. [Online]. Available: http://parnec.nuaa.edu.cn/xtan/data/ClosedEyeDatabases.html

MKHUSELI NGXANDE received the B.Sc. and M.Sc. degrees in computer science from the University of Fort Hare, South Africa. He is currently pursuing the Ph.D. degree with the University of KwaZulu-Natal. His research interests include computer vision, image processing, deep learning, and high-performance computing. He has been awarded a Ph.D. studentship with the Council for Scientific and Industrial Research, South Africa.

JULES-RAYMOND TAPAMO (Member, IEEE) received the Ph.D. degree in computer science from the University of Rouen. He is currently a Professor of computer science and engineering with the School of Engineering, University of KwaZulu-Natal, South Africa. His research interests include image processing, computer vision, machine learning, biometrics, intelligent monitoring, activity recognition, and data science. He is a member of the IEEE Computer Society, the IEEE Signal Processing Society, the IEEE Geoscience and Remote Sensing Society, the IEEE Computational Intelligence Society, and the ACM.

MICHAEL BURKE (Member, IEEE) received the bachelor's degree in electronic engineering from the University of Pretoria, the master's degree in electronic engineering from Stellenbosch University, and the Ph.D. degree in statistical signal processing from the University of Cambridge. He is currently a Research Associate with the University of Edinburgh. Prior to this, he led the Mobile Intelligent Autonomous Systems Group, Modeling and Digital Science Unit, Council for Scientific and Industrial Research, South Africa. He holds a Visiting Lecturer position at the University of Witwatersrand. His research interests include robot learning, and applications of computer vision and machine learning to robotics.